情報科学
【AI・データサイエンス】

第1回
様々なデータとデータ分析の基本

データとは何か？なぜ学ぶ必要があるのか？
データ分析の基本：①予測，②傾向や関連の発見，③分類・グルーピング

データとは何か？
なぜ学ぶ必要があるのか？

データとは

「データ」とは？ (デジタル大辞泉より)

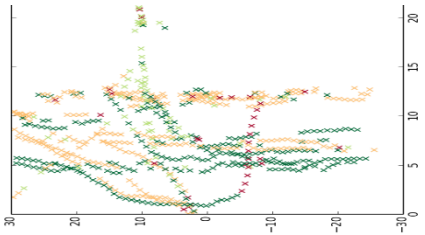
1. 物事の推論の基礎となる事実。また、参考となる資料・情報。「―を集める」「確実な―」

2. コンピューターで、プログラムを使った処理の対象となる記号化・数字化された資料。

本講義では主に2.を扱う

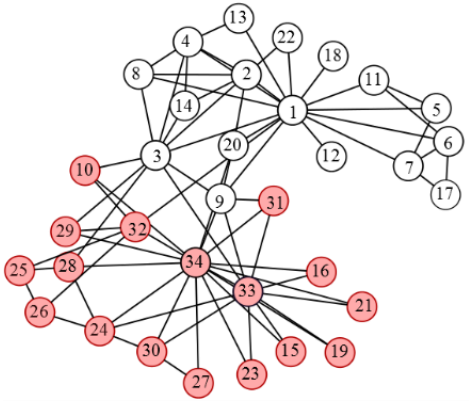
データとは？

伊都キャンパス内センサで計測した人流データ



- 測定値
  - 体温，体重，消費カロリー，人流
- メディアデータ
  - 画像(次スライド)，動画像（ビデオ），音声
- ラベルデータ
  - 患者の病名，地点名・駅名，生物種
- ネットワーク(関係データ)
  - 空手クラブメンバーの仲良し関係

シロイヌナズナ by Alberto Salguero@Wikipedia

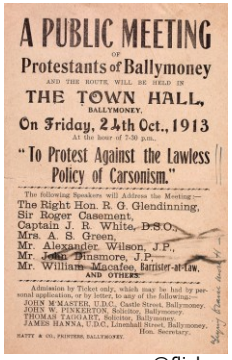


Zachary's Karate Club by Cuneytgurcan @Wikipedia

個々の○がメンバー1人に対応．
仲良しは線で結ばれる

メディアデータの代表例：画像

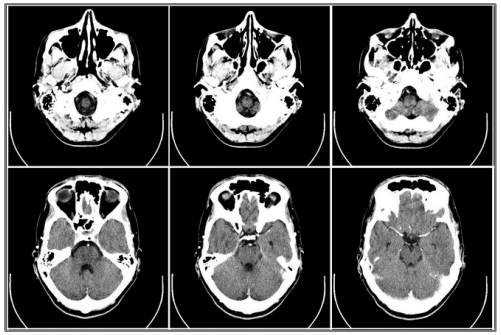
- カメラ画像
- 文字，文書，記号，標識，ナンバープレート
- 顔，指紋，虹彩，耳，唇，掌の静脈
- CT・MRI・X線などの医用画像



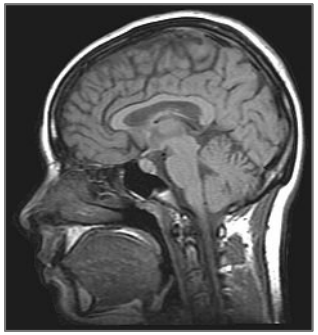
commons@flickr

IAM face dataset

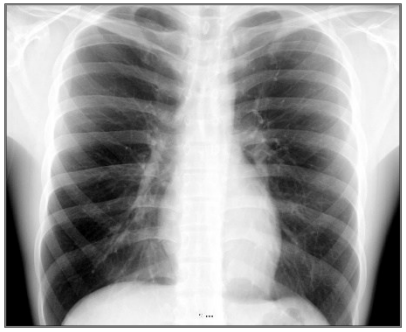
@wikipedia



CT画像@wikipedia



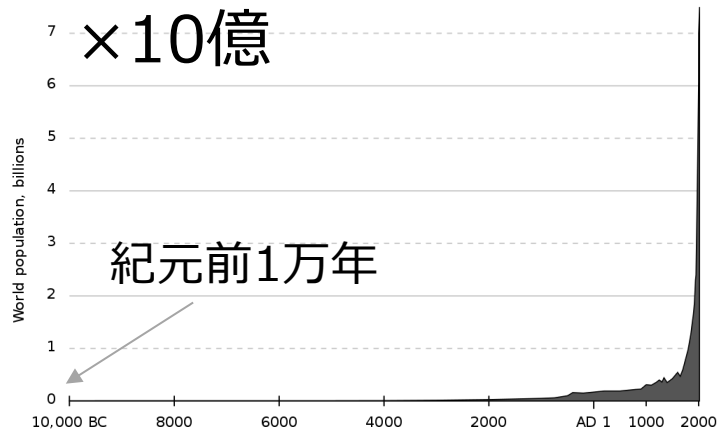
MRI画像@wikipedia



X線画像@wikipedia

データの種類～別の角度から：前後関係のあるデータ＝「系列データ」

- 時々刻々と得られる系列データ（時系列データ）
  - 動画像＝静止画像の時系列
  - 行動，ジェスチャ，歩行，ゲーム操作
  - 音声信号．対話系列
  - 心拍数変化，呼気量変化
  - PM2.5濃度変化，気温変化
  - 人口推移



- 時間とは関係のない系列データ
  - 文字列（文章）
  - DNA系列

cgcacagtgg atcctaggcg ttactaggtc
ttcaattctt gaactaattg ttttcgggtt ...

データの一般的な4分類 (1/3)
同じ数字に見えても，実は性質が違うことがある

- 量的データ
  - 比率データ
    - 積や除算ができる．和や差もできる．Ex. 体重．年収．長さ
  - 間隔データ
    - 積や除算に意味がない．ただし和や差はできる．Ex. (華氏・摂氏で測る)温度，西暦年

「温度が２倍」「温度70%減」とは言わない

- 質的データ
  - 順位データ
    - 四則演算（加減乗除）すべて意味がない．ただし並べることはできる．
    - Ex. アンケート結果（５:非常によい，4:よい，3:ふつう，2:わるい，1:非常に悪い)．成績順
  - カテゴリデータ
    - 形式的に数字になっているだけ．
    - Ex.「1:女性, 2:男性」，電話番号，背番号，バスの系統番号

「非常によい-ふつう=わるい」とはならない

データの一般的な4分類 (2/3)
比例データと間隔データの違いをもう少し

- 見分け方① ゼロが絶対的か相対的か
  - 気温は間隔データ．摂氏0度は人間が適当に決めたもの(＝相対的)なので
  - 標高も間隔データ．現在の標高0mは人間が決めたもの

  - 絶対温度は比例データ．絶対0度は全ての運動がゼロになるので
  - 重さや長さは比例データ．ゼロは本当にゼロ（何もない状態）なので

- 見分け方② 比に意味があるか？
  - 西暦は間隔データ．　西暦1500年は，西暦1000年の1.5倍ではない
  - 気温は間隔データ．気温30度は，気温20度の1.5倍ではない

  - 重さは比例データ．30グラムは20グラムの1.5倍
  - 収入は比例データ．給料40万は20万の2倍

データの一般的な4分類(3/3)
表としてまとめると…

	名称	可能な演算	主な代表値	主な事例
量的データ	比率データ	＋ − × ÷	各種平均	質量，長さ，年齢，時間，金額
	間隔データ	＋ −	算術平均	温度（摂氏），知能指数
質的データ	順位データ	＞ ＝	中央値，最頻値	満足度，選好度，硬度
	カテゴリデータ	度数カウント	最頻値	電話番号，性別，血液型

データの種類によって使える手法が
大きく異なってくる

別の角度からの分類：構造化データと非構造化データ

- 構造化データ
  - 簡単に言えば，表形式のデータ
  - 例：「全県の毎月の平均降水量」の表 →

1-8　降水量（平年値）（昭和56年～平成22年）

観測地点	降水量（mm） 年計	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月
札幌	1,107	114	94	78	57	53	47	81	124	135	109	104	112
青森	1,300	145	111	70	63	81	76	117	123	123	104	138	151
盛岡	1,266	53	49	81	88	103	110	186	184	160	93	90	71
高知	2,548	59	106	190	244	292	346	328	283	350	166	125	58
福岡	1,612	68	72	113	117	143	255	278	172	178	74	85	60
佐賀	1,870	57	78	129	156	198	339	339	197	180	76	76	48
長崎	1,858	64	86	132	151	179	315	314	195	189	86	86	61
熊本	1,986	60	83	138	146	196	405	401	174	170	79	81	54
大分	1,645	45	65	112	129	150	274	253	172	220	121	69	34
宮崎	2,509	64	91	182	213	239	429	309	290	355	182	95	60
鹿児島	2,266	78	112	180	205	221	452	319	223	211	102	92	71
那覇	2,041	107	120	161	166	232	247	141	241	261	153	110	103

資料　気象庁「2010年平年値」

- 非構造化データ
  - 文章，画像，音がその代表例
  - 「表形式」にはならないので「非構造化データ」と呼ばれる
  - スマートフォンやパソコンで日々読んだり見たり聞いたりしているが，これらもデータ

非構造化データについては，そのうち触れます

あらゆる分野で
データ分析の必要性が
高まっている

データの分析
＝データから意味のある情報を引き出す

- 直感的には
  - データ＝コーヒー豆
  - 分析結果＝(おいしい)コーヒー

焙煎→粉砕→湯による成分抽出

- 適切な分析方法を用いなければ，意味のある情報は抽出できない
  - コーヒー豆を炒めて食べても，おいしくない

なぜ「今」データ分析なのか？ 3つの大きな理由

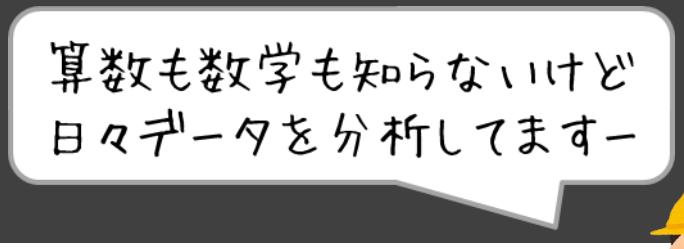
- 学術的・社会的要請
  - 様々なことを説明・説得するためには，その根拠が必要
  - さらに根拠は「主観」ではなく「客観的な数値」で表現されるべき
- データからの要請
  - データが大規模・複雑・多様に
  - そのため，手計算での分析は無理．計算機で分析する必要
- データ分析技術の進展＝以前はできなかった分析ができるように
  - 計算機の能力が向上
  - 数値分析法，機械学習(特に深層学習）の進歩
  - オープンソース化，無料ライブラリ，技術解説サイト

皆さんの先輩も様々なデータ分析を行っている

「色々あるんだ！自分の学部とも関係あるんだ！」ということがわかれば十分

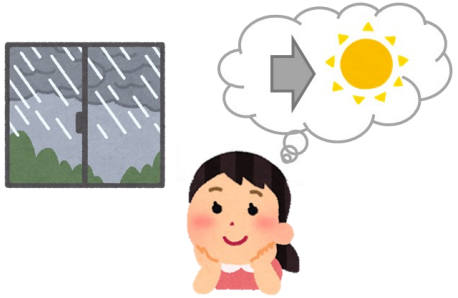
- 外部刺激応答の理解に向けたインターフェイスとしての**脳オルガノイド回路**の開発
- パルス加熱駆動型ガスセンサの応答波形解析による**微量 VOC ガス**の高精度識別
- Designing Ionic Liquids with Generative AI for **Carbon Neutral** Technology
- **芸術文化データ**の進化解析のための統計モデリングの研究と教育
- ベイズ計測導入による**強相関希土類化合物**の高精度電子状態測定技術の開拓
- 感覚精度の個人差に対してロバストな**心理物理実験**手法の提案と検証
- 公的ミクロデータを活用した**貧困状況**実態把握のための解析
- **疲労感**を推定する有効な行動指標の検討および推定モデルの構築
- Mining **Students' Learning** Strategy and Predicting Knowledge Gain in Reading using Learning Logging, Eye-tracking and Physiological Measures
- **クルマエビ養殖池**水底環境の見える化による養殖生産支援
- ハイエントロピー**酸化物**の構造決定の効率化
- Machine Learning to improve and optimize parameter prediction in High Temperature **Gas Cooled Reactor**
- 質量分析シグナルから**食品品質**を予測・生成するフードミクス AI の基盤構築
- ベイズ統計を用いた **Grad-Shafranov 方程式**の解空間の探索
- Intelligent Control of Complex Renewable **Energy Systems**: Model Predictive Control and Reinforcement Learning Combined with Deep Learning Prediction
- 溶接条件をパラメータとした深層学習による**自動溶接**・積層造形検査技術の構築
- 視覚カテゴリーの素早い認識を実現する脳内基盤:**脳波データ**解析による検討
- Multi-Aspect Assessment of Student Reflections to Enhance **Personalized Learning**
- 「機械学習」「**反応**座標抽出」「反応・状態遷移」

研究者だけじゃない！
誰もが「無意識に」
データを分析しながら
生きている

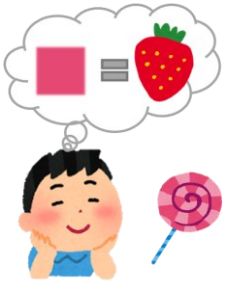


データ分析，主な３つのタスク：
難しそう？ いえいえ，皆さんも常にやってますよ

- 予測
  - 未来を予測することはできないか？



- （傾向や関連の）発見
  - これまでに気づかなかった傾向や関連などを発見できないか？

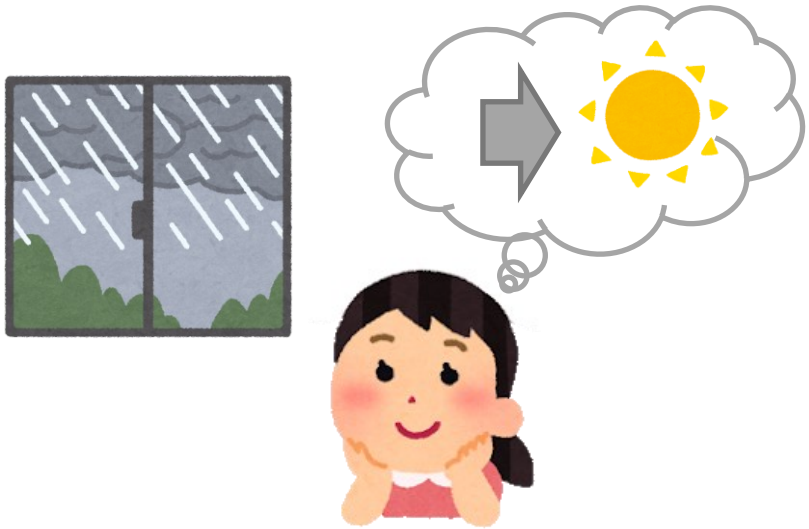


- 分類・グルーピング
  - 様々なデータを種類ごとに区別できないか？
  - たくさんのデータを「似たデータ」ごとにまとめられないか？

次スライドから，皆さんも常にやってることを説明

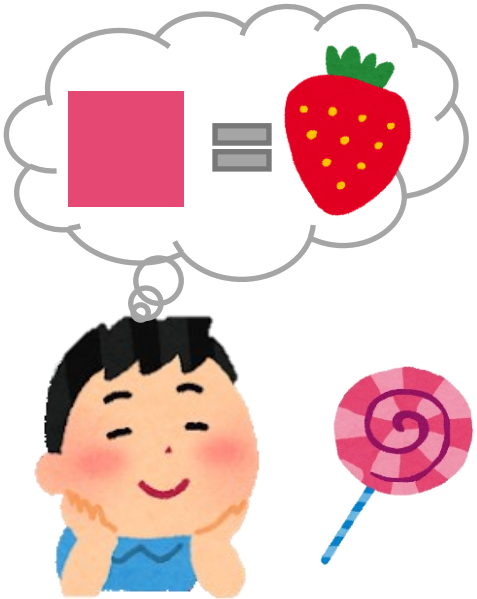
皆さんもやっているデータ分析：予測

- 「ここのところずっと雨なので，明日は晴れるだろう」
- 「あと2 時間もすれば，この宿題も終わるだろう」
- 「次はカーブを投げてくるだろう」
- 「これだけ勉強すれば，100点取れるだろう」



皆さんもやっているデータ分析：傾向や関連の発見

- 「赤いアメはイチゴ味」
- 「いい子にしていればサンタがプレゼントを持ってくる」
- 「夜更かしすると翌朝起きられない」



皆さんもやっているデータ分析：分類・グルーピング

- 「目の前の動物は犬か猫か」
- 「母親の表情がいつもとちょっと違う」

- 「自分が好きな本と嫌いな本がある」

データ分析は，数学が苦手な人どころか，算数を習ってすらいない幼児にとっても，極めて身近なもの

- 先入観は捨てよう
  - 「データ＝数字が並んだ無味乾燥なもの」→ No!
  - 「データ分析＝難しくて専門家しかできない」→ No!
  - 「自分の人生には関係ない」 → No!

- その面白さを是非理解してほしい
  - ある意味「柔らかく」「結果も色々ありうる」人間らしい話

- 自分自身が日々（無意識に）どのようなデータ分析をしながら生きているのかを考えてみても，きっと楽しいはず！

データ分析の基本：
①予測
②傾向や関連の発見
③分類・グルーピング

３つの分析タスクについて，もう少しだけ詳しく

データ分析の基本①
予測

みんなやってる予測．でも難しい．

身近な予測①
「未来」を予測するケース

- 試験の勉強
  - 過去の傾向からみて，明日はきっとこの問題が出るだろう

- スポーツ
  - 次はストレートを投げてくるに違いない

- 買い物
  - この値段・素材のものを買えば，5年は大丈夫だろう

- 天気予報
  - 過去の天気データを用いて，明日以降の天気を予測

他にも…
- 株価の予測
- 競馬等のギャンブル
- 就職活動

などなど

身近な予測②
予測は未来だけとは限らない．「だろう」がつけば予測

- 画像認識
  - (無意識に)「この動物は犬だろう」
  - (無意識に)「あ，機嫌が悪そうだ」
  - この本（表紙とタイトル）は，きっと面白いだろう

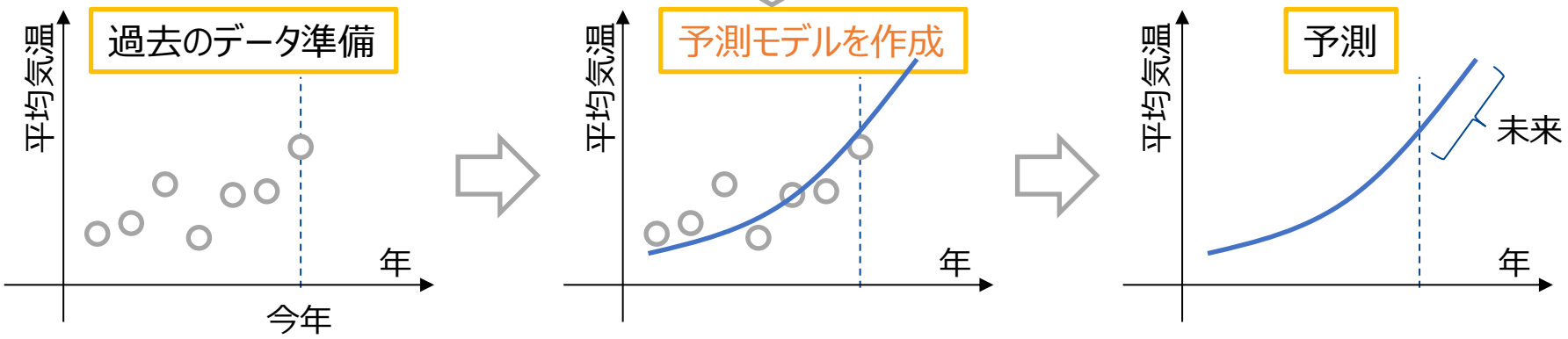
- 推量・診断
  - これぐらい勉強すれば，これぐらいの点数は取れるだろう
  - この体温ならば，インフルエンザだろう

- 因果推論(＝こういう結果になったのはこういう原因があったからだ)
  - 警察の推理，故障原因の推定，考古学

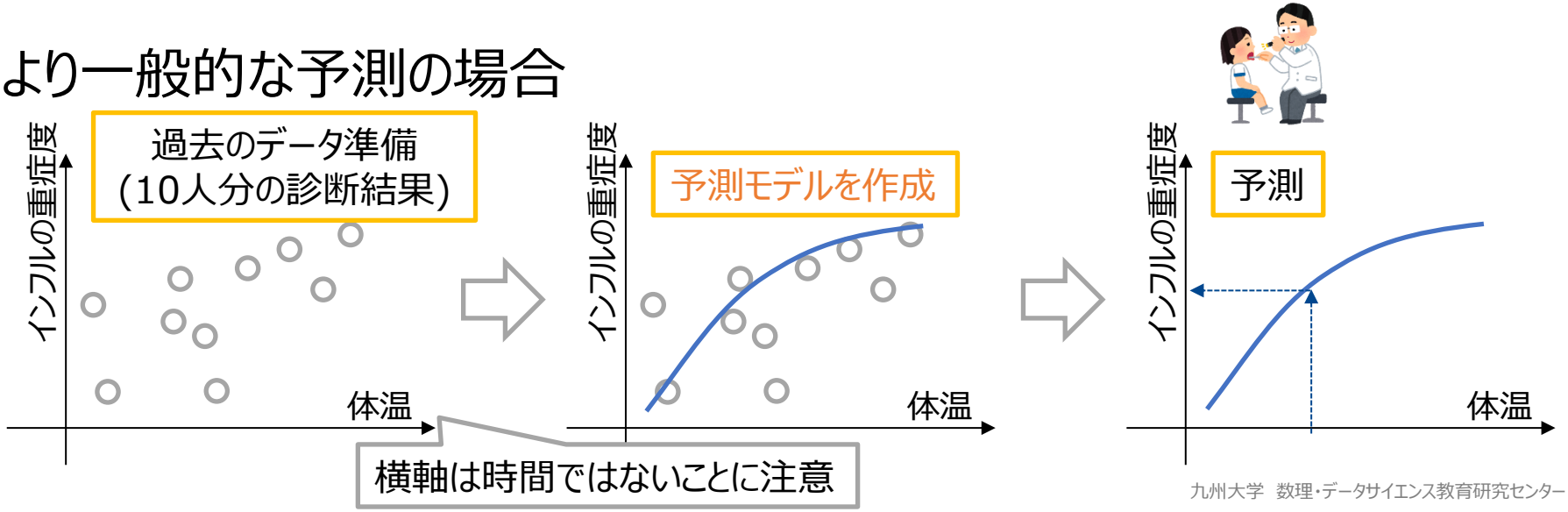
- 推薦
  - このユーザーなら，この商品なら買ってくれそう！

データを用いた予測の方法：
難しそうに感じるかもしれませんが，みんな無意識にやっていることです

- 未来を予測する場合

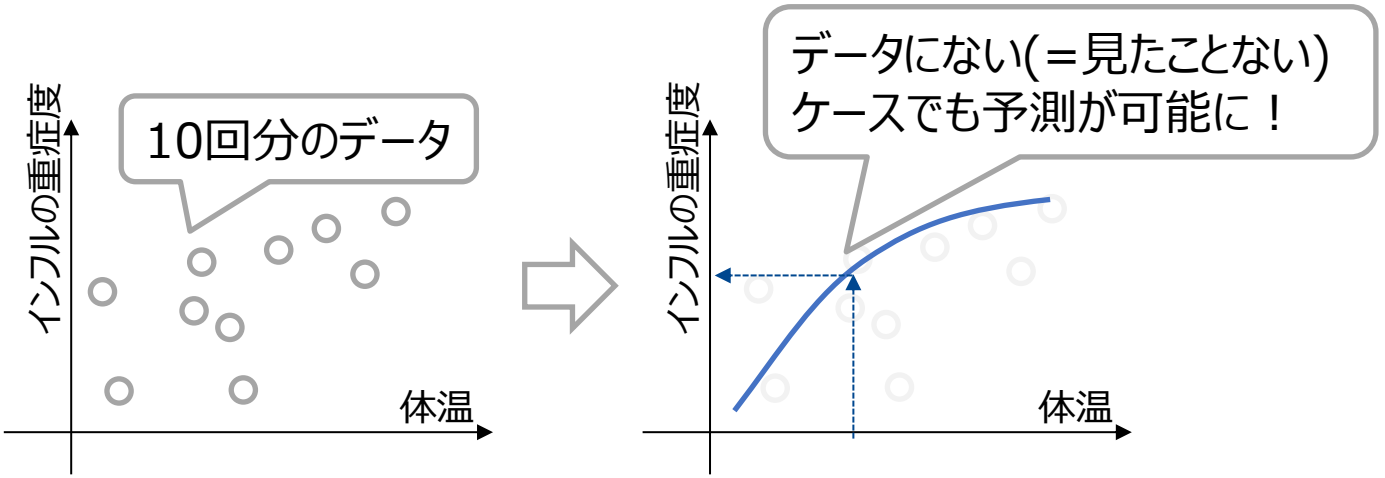


- より一般的な予測の場合



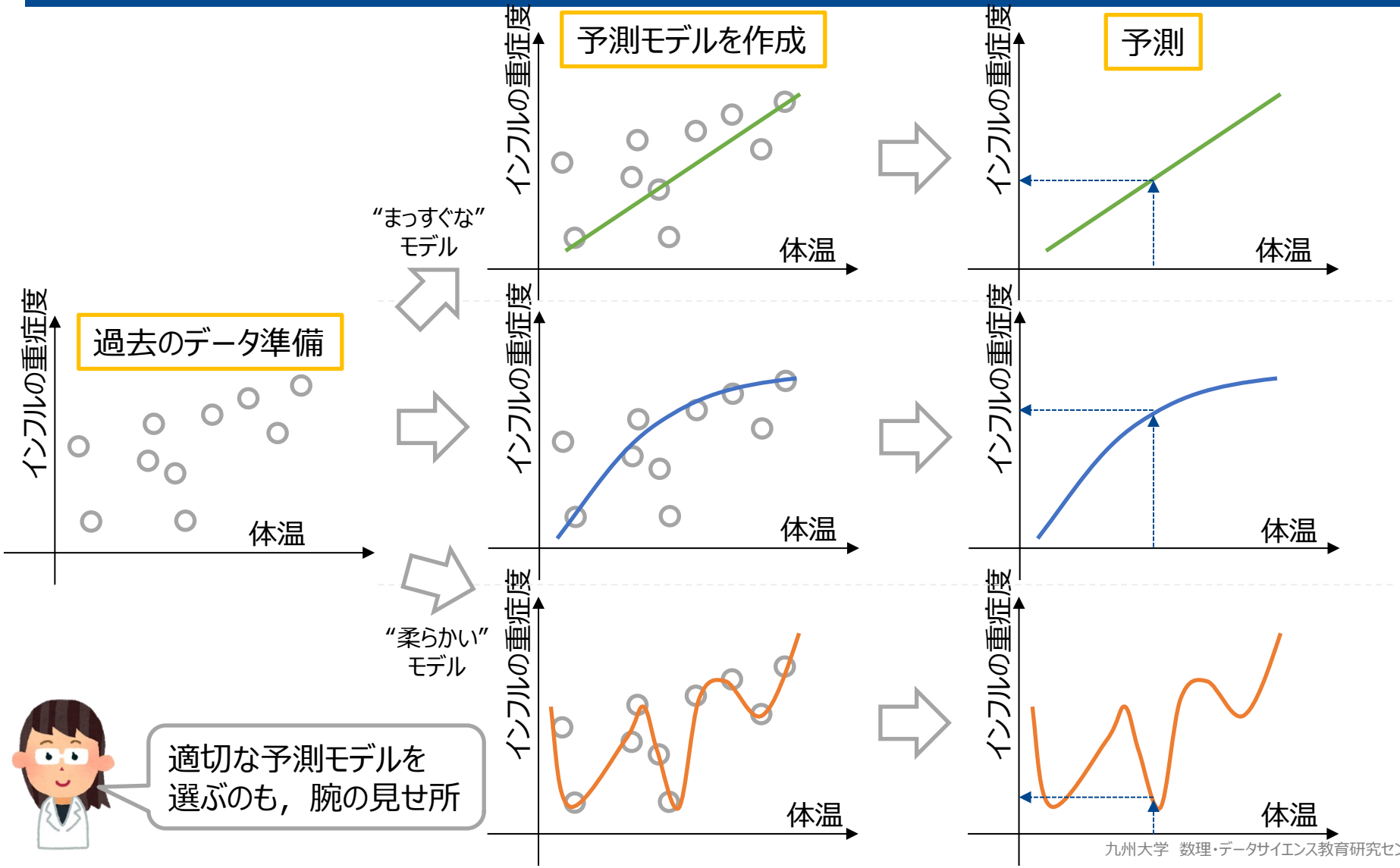
予測モデルができると，うれしいところ

- 過去になかった状況に対しても予測可能



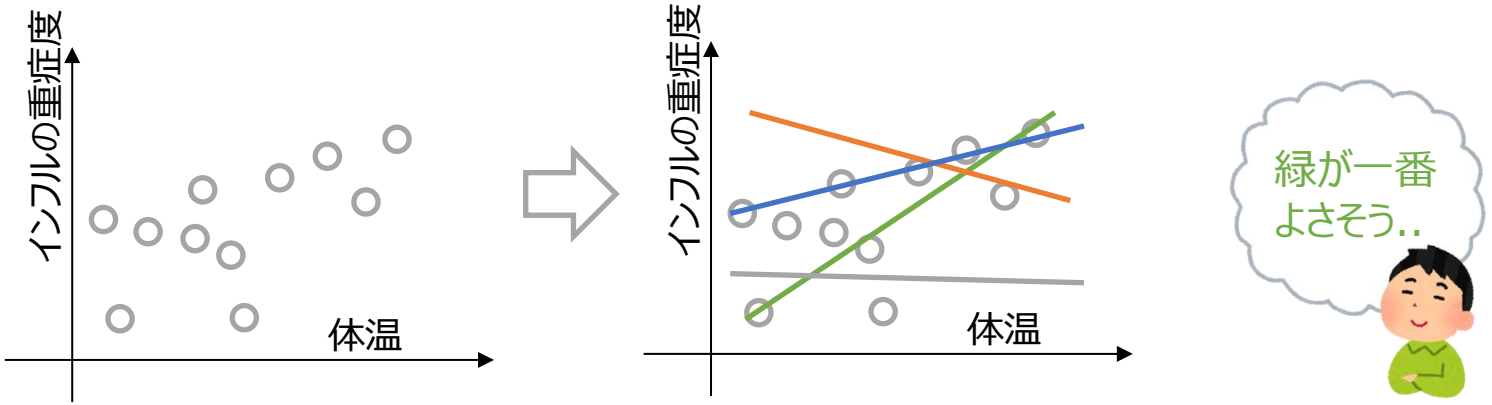
- 我々人間も，過去の数回の経験(＝データ)に基づいて，
  未知の状況でも何らかの予測を行いながら生きてる！
  - 例：医師も，他の患者の診察結果に基づいて，初診の患者を診察している
  - 例：見たことのないタイプの犬でも，過去に見た犬に基づき，犬とわかる

予測モデルは様々考えられる．そして，モデルによって予測結果は異なる(精度が違う)



さらに：
同じ予測モデルでも，「あてはめ方」は色々

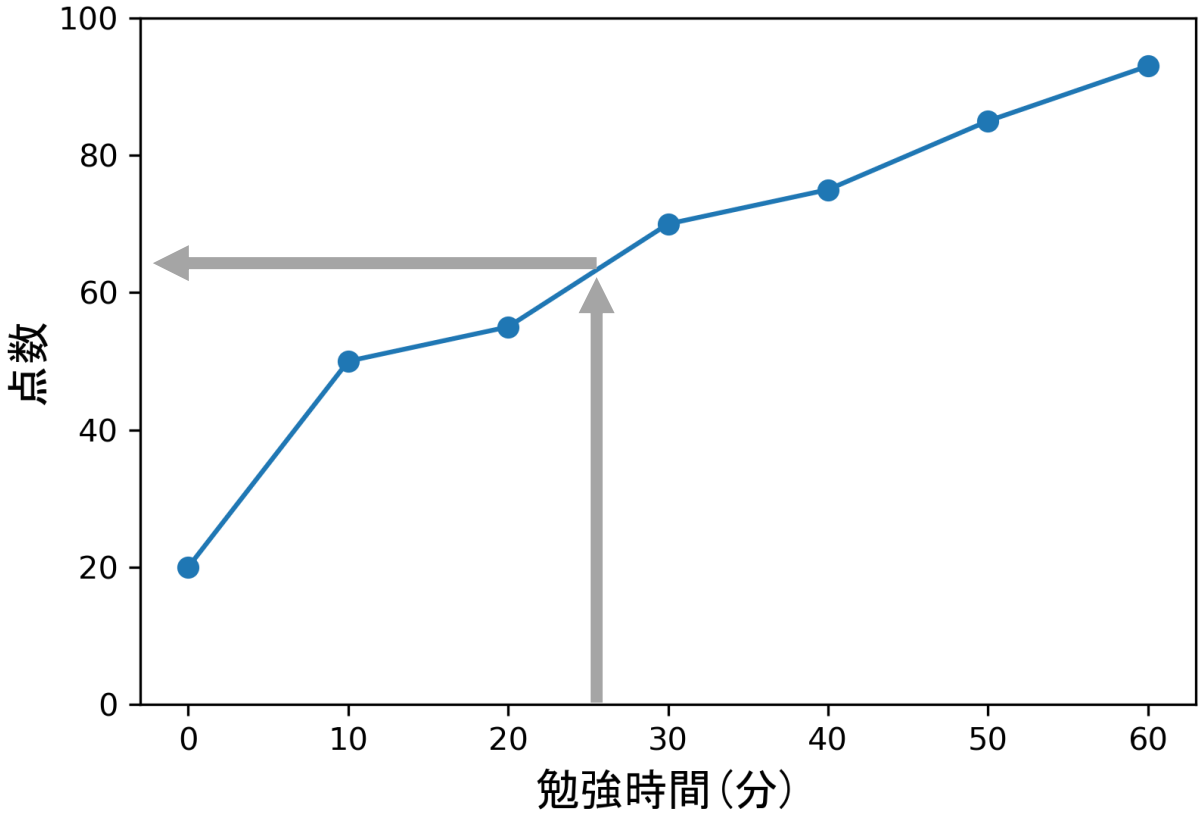
- 例えば同じ「まっすぐな予測モデル(線形予測モデル)」でもデータへの「あてはめ方」は色々考えられる



- なるべく多くのデータを正確に予測できるように，適切にあてはめる必要がある
  - これもまた「腕の見せ所」
  - この点をより深く知りたい人は → 最小二乗法

小学生でも予測をやっている！ 折れ線グラフ，実は最も単純な「予測」

- 予測モデル＝折れ線

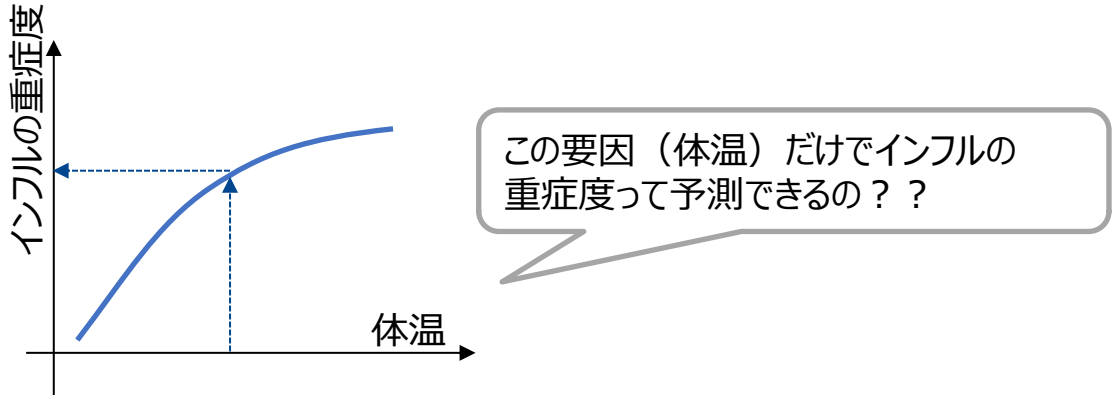


- 測定していない勉強時間(ex. 25分)でも点数を予想可能

それでも予測は難しい～4つの理由 (1/2)

1. 過去のデータを十分に集められない場合がある
   - 「あなたの10年後の給料」を予測するためには，「あなたと似たような人」をたくさん集める必要がある

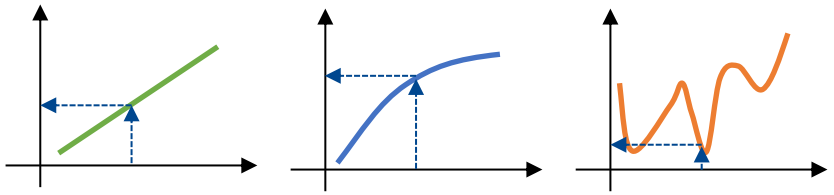
2. 予測結果を決める要因が不明な場合がある
   - インフルの重症度は体温だけでよいのか？
   - 上記の「10年後の給料」予測に必要な要因は？
   - 天気予報のように要因がほとんど無限に存在する場合もあり



それでも予測は難しい～4つの理由 (2/2)

3. 現時点と予測時点では状況が違う場合がある
   - 2年後に突然不況が起こったら，「10年後の給料」予測結果は外れる
   - ＝いつまでも同じ予測モデルが使えるとは限らない

4. どの予測モデルを使えばよいかは，自明ではない



どのモデルを
選ぶべき？？

… 上記4つ以外にも様々な難しさがある！

専門家でも，
予測は難しい

データ分析の基本②
傾向や関連の発見

「背が高いと数学力も高い」なんて話にだまされないように

発見とは？

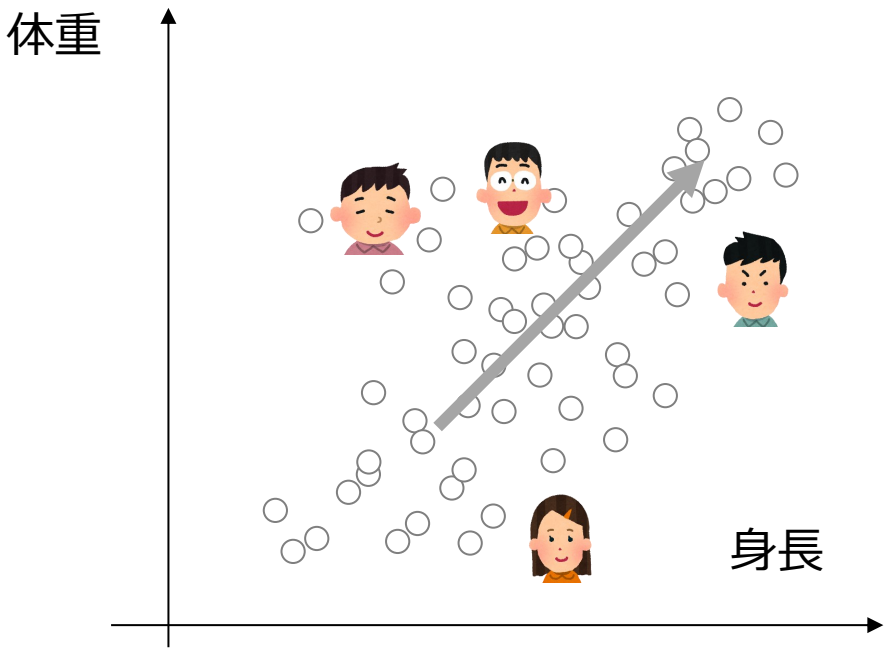
- 大規模なデータの中に潜む傾向を見つける方法

- 例：
  - 商店の販売データを大量に準備
  - データから「商品A を買う人は商品B も買う」傾向を発見！
  - 商品A の横にB も並べて陳列すれば売り上げが伸びるかも！？

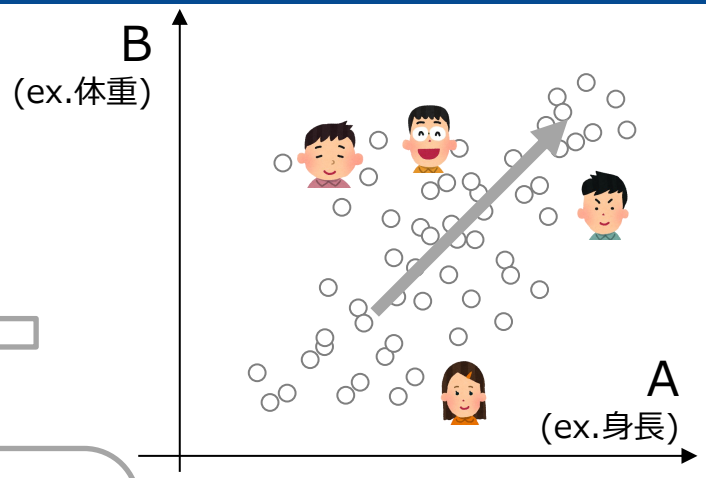
- 「発見」のための代表的手法
  - 相関(correlation) 分析
  - 頻出パターン発見(frequent pattern discovery)

相関分析：
そもそも相関とは？

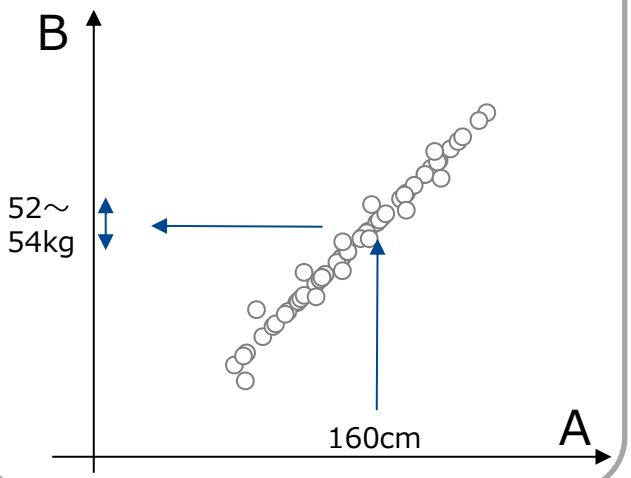
- 身長Aと体重Bのように，「Aが増えればBも増える」というような傾向があるとき，「AとBは相関する」という



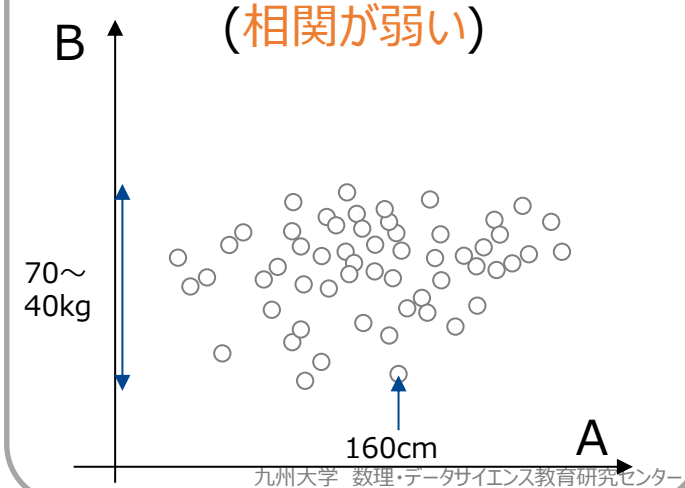
相関分析：
相関には「強さ」がある＝傾向には強さがある



傾向が強い
＝Aをわかるとも結構わかる
(相関が強い)

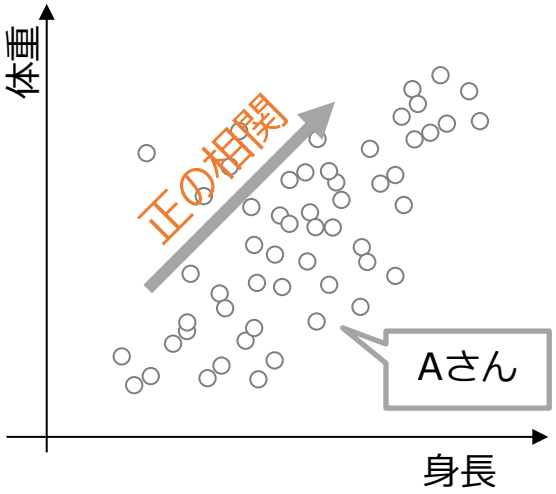


傾向が弱い
＝Aがわかっても
Bを知るのにはあまり役に立たない
(相関が弱い)

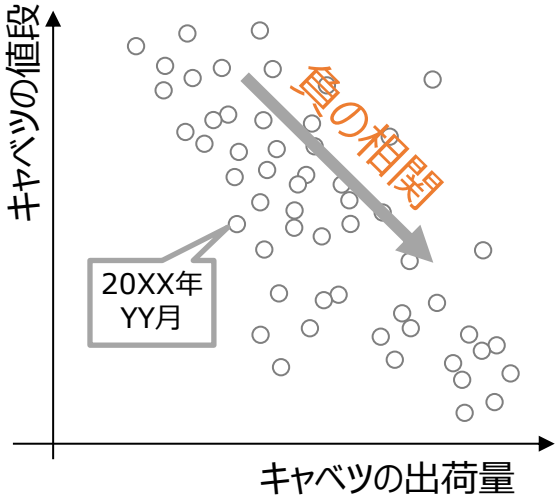


相関分析：
相関には「正の相関・負の相関・無相関」がある

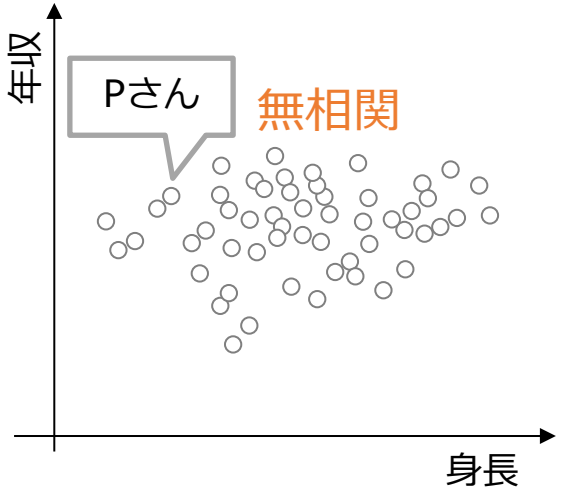
身長が高いほど体重が重い傾向
（正の相関が強い）



出荷量が多いほど価格が下がる傾向
（負の相関が強い）

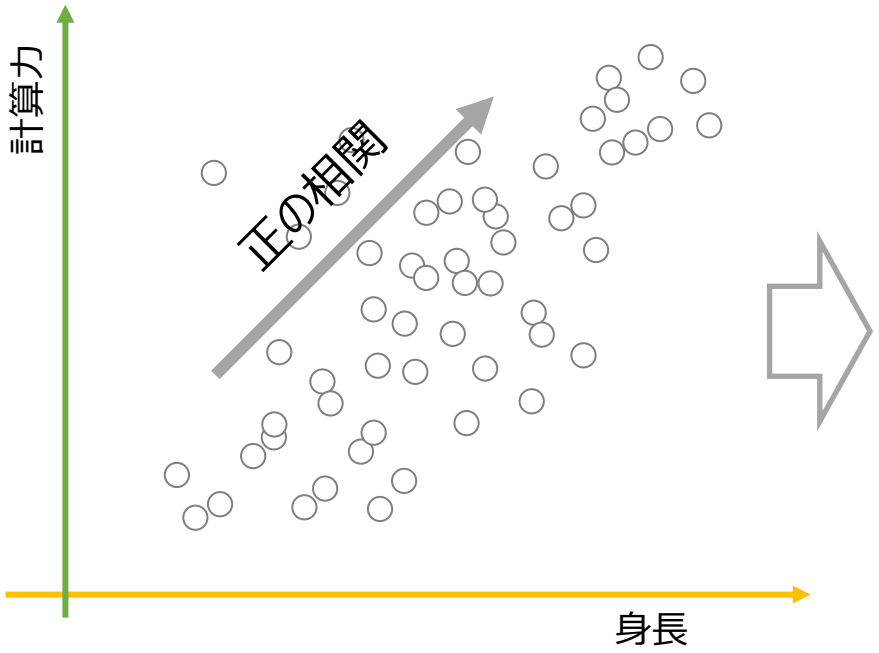


身長と年収の間には特別の傾向は見られない
（相関ゼロ＝最も相関が弱い）

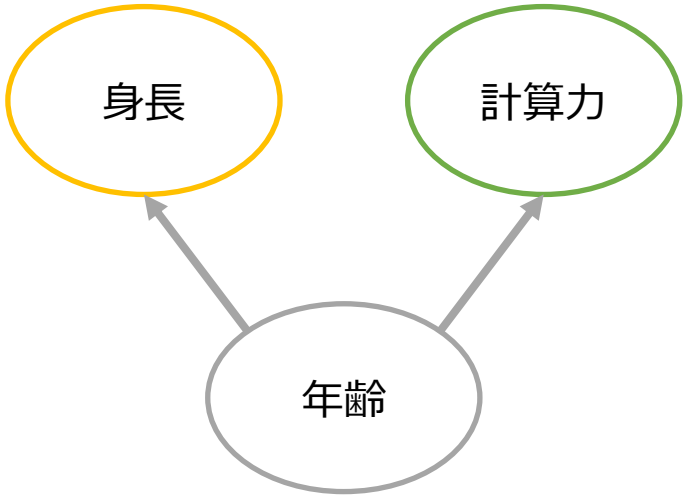


- 考えてみよう
  - 「勉強時間」と「テストの点数」は {正の相関，負の相関，無相関}？
  - 動画の「長さ」と「データ量」は {正の相関，負の相関，無相関}？
  - カレーライスの「分量」と「値段」は {正の相関，負の相関，無相関}？？？

擬似相関には気を付けよう！
背が高いと算数が得意!?



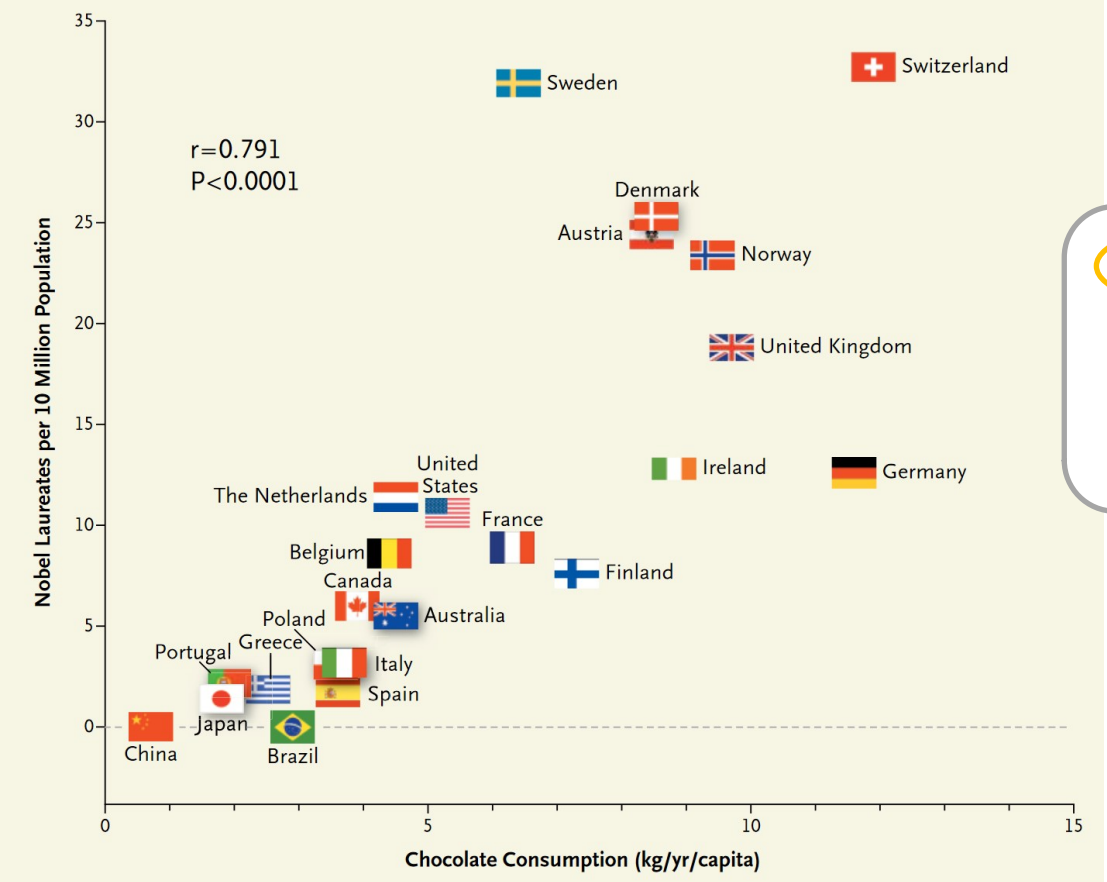
「身長」と「計算力」には
正の相関があった！
（背が高いほうが計算得意！）



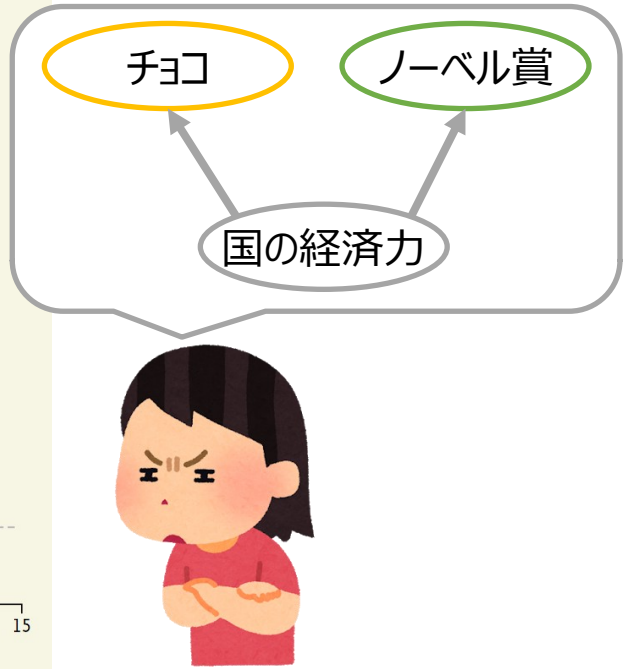
「年齢」という要因のために
「見かけ上」相関しているだけ

擬似相関には気を付けよう！
チョコを食べる国はノーベル賞が多い!?

人口1千万人あたりのノーベル賞受賞者



チョコレート消費量



[Messerli, Chocolate Consumption, Cognitive Function, and Nobel Laureates, The New England Journal of Medicine, 2012]

「チョコを食べる国はノーベル賞が多い」ことを本気で証明するのは，結構大変

1. 同じような（年齢，健康状態，食生活，住所，成長過程などが似た）人々を集め，ランダムに2群に分ける

2. 第一群はチョコレートを食べる，第二群は食べないという条件以外は，極力同じような状況で過ごしてもらう.

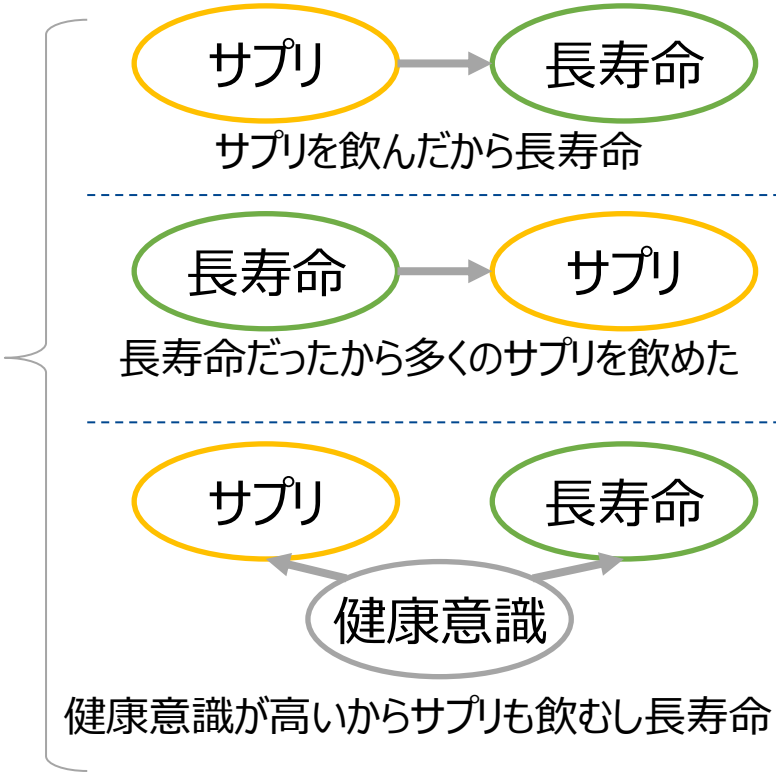
3. 数年後に，2 群の間で，頭脳に差が出るかどうかをテストする（集めた人々のノーベル賞の数で評価してもよい）

「相関と因果関係は違う」ことにも気を付けよう！

- 因果関係＝「こういう原因だから，こういう結果になった」
- AとBに相関関係があっても，AとBのどちらが原因・結果かは不明
- さらには擬似相関の可能性もあるので要注意

「サプリ摂取量」と「寿命」には
正の相関があった！

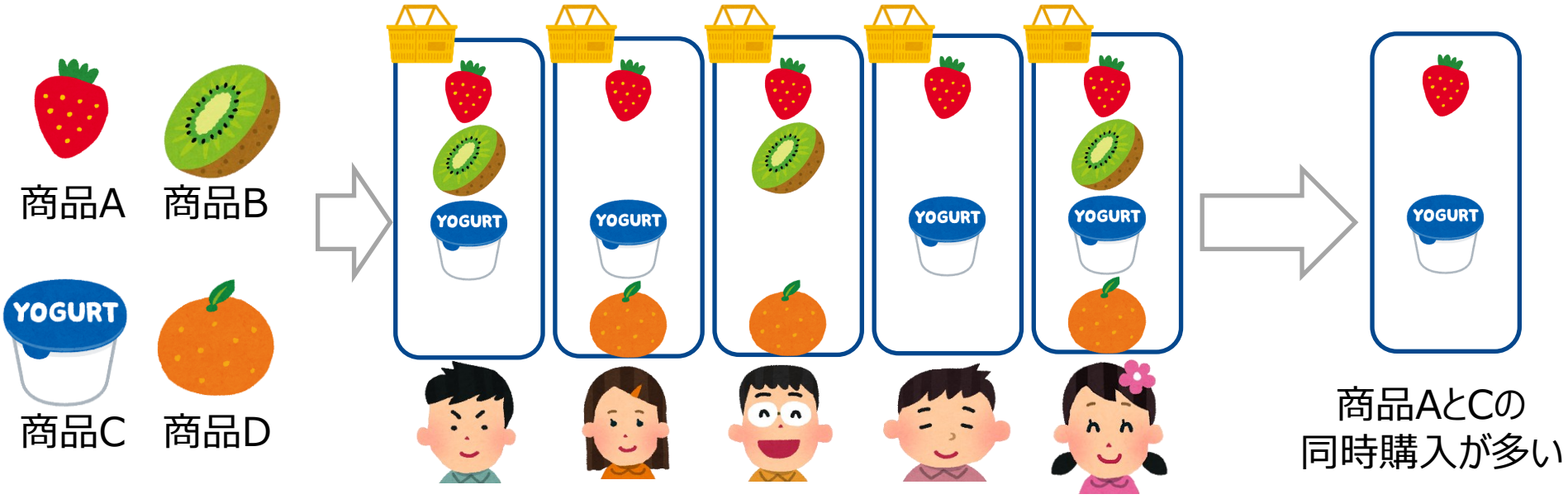
としても…



どれが本当かは不明

因果関係を正しく知りたい人は→付録

頻出パターン発見とは？
バスケット解析



- 「ヨーグルトとイチゴを近くの売り場に置く」と売り上げアップ？
- アンケート結果の分析にも応用可能
  - 例：「Q2をyesの人はQ5もyes」が多い，とか

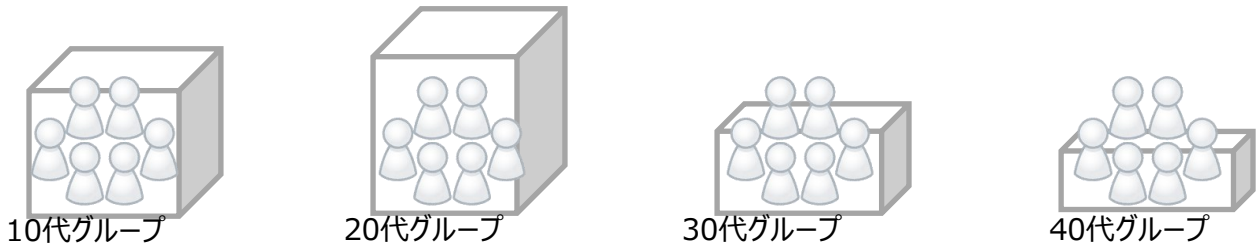
データ分析の基本③
分類・グルーピング

「データが似ている」とは，どういうことか？

グルーピングとは何か？

- データをいくつかのグループにわけること
  - グループ単位で見ることで，データ全体の状況把握が容易に！

- 例えば，国民全員分の年齢データ
  - 10代，20代，…とグループにわければ，その国にどれぐらいの年齢の人が多いか，ざくっと把握可能

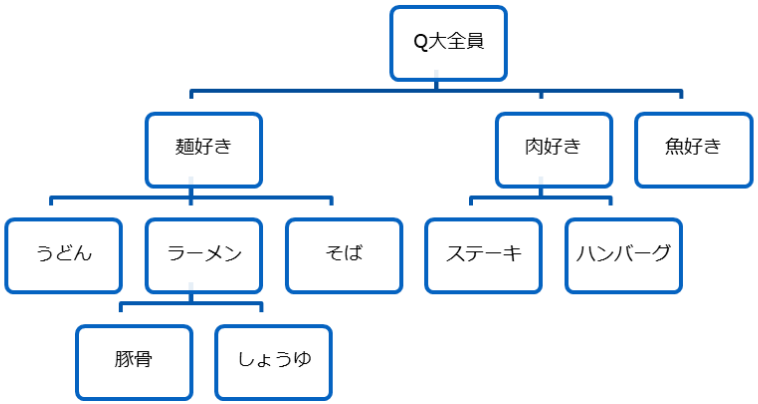


グループがあらかじめ決まっている場合

- 例
  - 10代，20代，30代…
  - 麺好き，肉好き，魚好き…
- “組合せ”も可能
  - 「20代」で「魚好き」

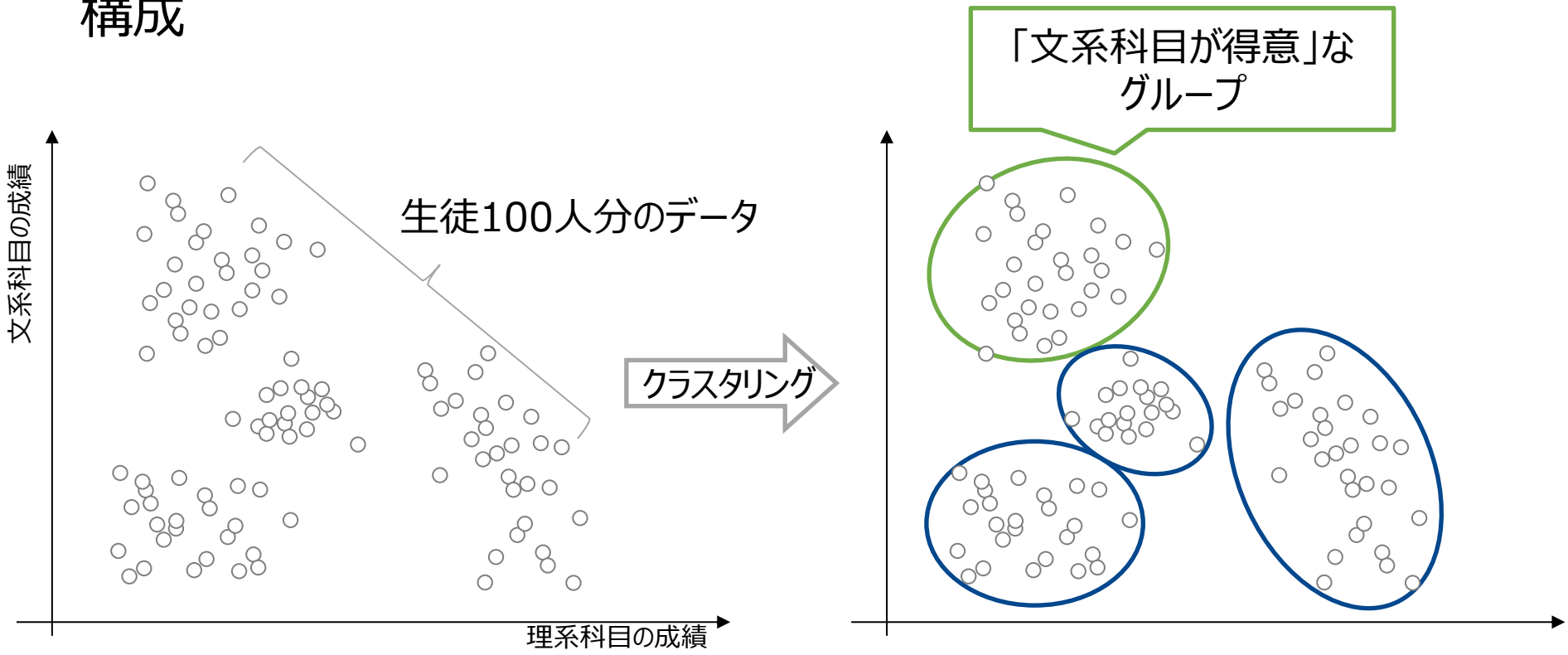
	10代	20代	30代
麺好き	10代の麺好きグループ	20代の麺好きグループ	30代の麺好きグループ
肉好き	10代の肉好きグループ	20代の肉好きグループ	30代の肉好きグループ
魚好き	10代の魚好きグループ	20代の魚好きグループ	30代の魚好きグループ

- “階層化”も可能



グループがあらかじめ決まっていない場合：クラスタリング

- 似たデータが同じグループになるようにすることで，自動的にグループを構成

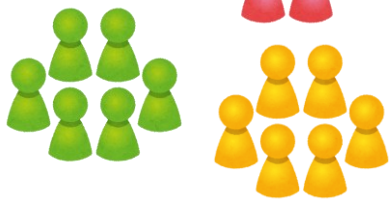


- クラスタ(cluster)：「似たデータのかたまり」＝「グループ」と同じ意味
  - 上の例では4つのクラスタに分けた

色々使えるクラスタリング

- 「起床・睡眠の時間」が似ていることで『人々』をグルーピング

- 「カスタマー層」が似ていることで『企業』をグルーピング

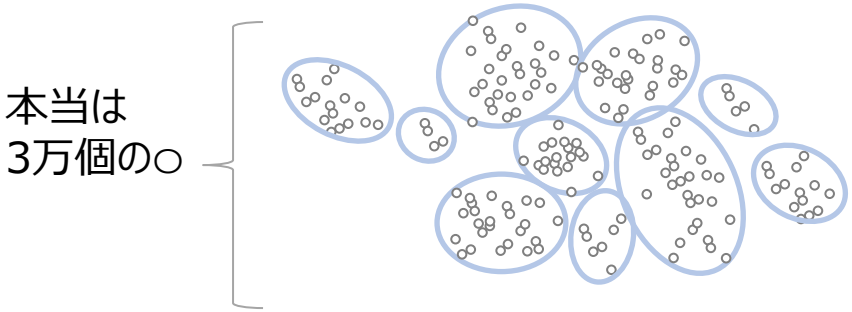


- 「産業構造」が似ていることで『国』をグルーピング
- 「味」が似ていることで『ラーメン』をグルーピング

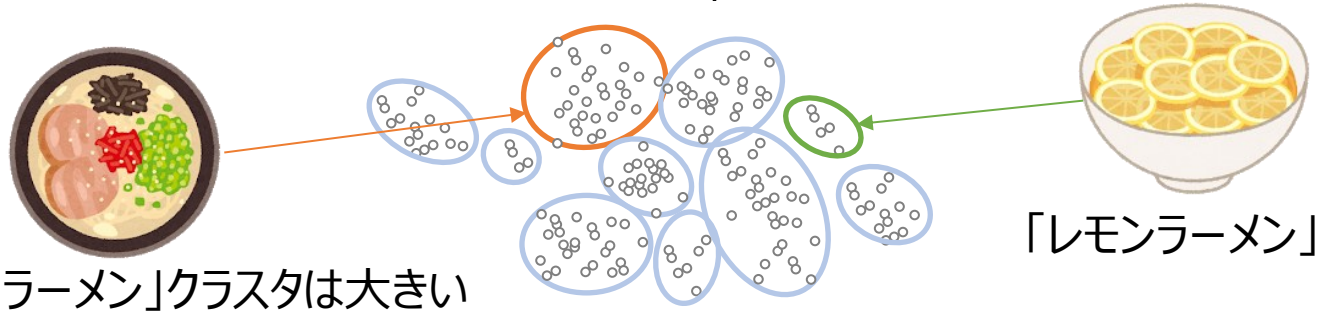
- 「曲調」が似ていることで『音楽』をグルーピング

クラスタリングの結果からわかること(1/2)：
全国約3万のラーメン店を味でクラスタリングできたら..?

- できたクラスタの数から，データ全体の多様性がわかる
  - 10クラスタなら，大きく分けて10タイプのラーメンが全国にある



- 各クラスタのデータ数で，各クラスタの勢力がわかる
  - たくさんのデータが含まれるクラスタは「メジャーな味」のラーメン
  - もしクラスタ10が全体の0.1%なら，それらは珍しいラーメンを提供

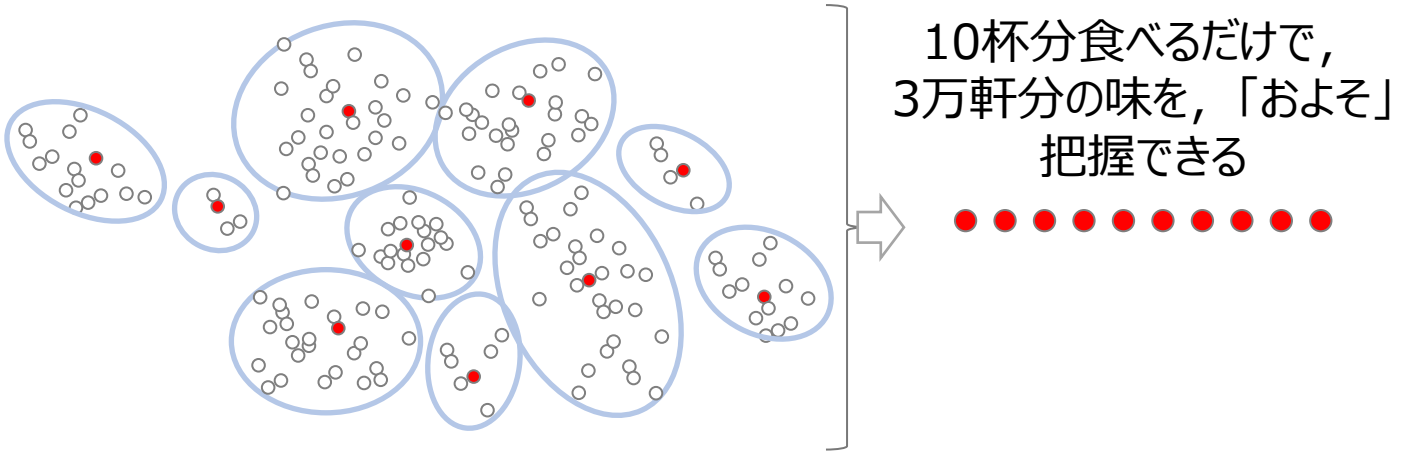


「豚骨ラーメン」クラスタは大きい

「レモンラーメン」クラスタは小さい

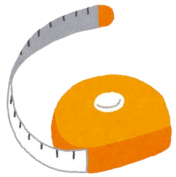
クラスタリングの結果からわかること(2/2)：
全国約3万のラーメン店を味でクラスタリングできたら..?

- 各グループの代表例を見ることで，全体を概観可能
  - 各クラスタから代表を選べば，我が国の10タイプのラーメンがわかる
  - 全3万店分を食べ歩くよりずっと効率的！



クラスタリングの実際：色々考えるべき点も多い

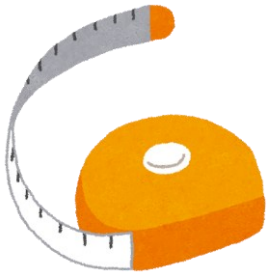
- クラスタリングは「似たデータを同じグループ」にする技術
- ではデータが「似ている」とはどういうことか？
  - 「どこが」似ている？
    - ラーメンの油っこさか，コクか，スープの色か，麺の固さか，トッピングの量か，など，どこに注目して似てる具合を測る!?
  - 「どれぐらい」似ている？
    - 麺の固さをどう測るか？　さらに2つの麺の固さの似ている具合はどう測るか？
    - さらに，脂っこさの似てる具合と，麺の固さの似てる具合を，同様に扱っていいのか？



この話，重要なので，何度か出てきます．
特に「ベクトル・距離・類似度」の項でもう少し深く説明します！

クラスタリングには「絶対的な正解」が存在しないことが多い

- 「似ている具合」の測り方について，数学的に「こうしなさい」とは決まっていない！
  - これが違えば，クラスタリング結果も当然変わってくる



- クラスタ数の決め方も，多種多様
  - 事前に決める方法もあれば，自動で決める方法も
  - 自動で決めるにしても，やはり何かしら基準が必要
    - 3万軒全部を１クラスタにしても，バラバラの3万クラスタにしても，「間違い」とか「法律違反」ではない
- 数学に基づいたデータ解析を使えば，「なんでもビシッと決まる！」というわけではない！
  - ある意味，数学ほど，自由なもの（＝どうとでもできるもの）はない
  - 高校数学を「答えがビシッと決まるから好き」と言ってた諸君，それは数学の一側面でしかない

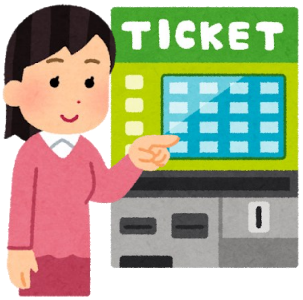
付録

データ分析の関連話題：最適化

最適化～人生で日々行ってること

- 最適化＝なるべくよい決定をしたり選択をしたりすること

- 人生は最適化の連続である
  - レストランで，何を食べるか決める
  - 次の一歩の足の位置
  - じゃんけん
  - 野球のピッチング・バッティング
  - 家に帰る道すじ
  - 今日一日のスケジュール
  - ライフイベント(進学や就職，結婚)，等々

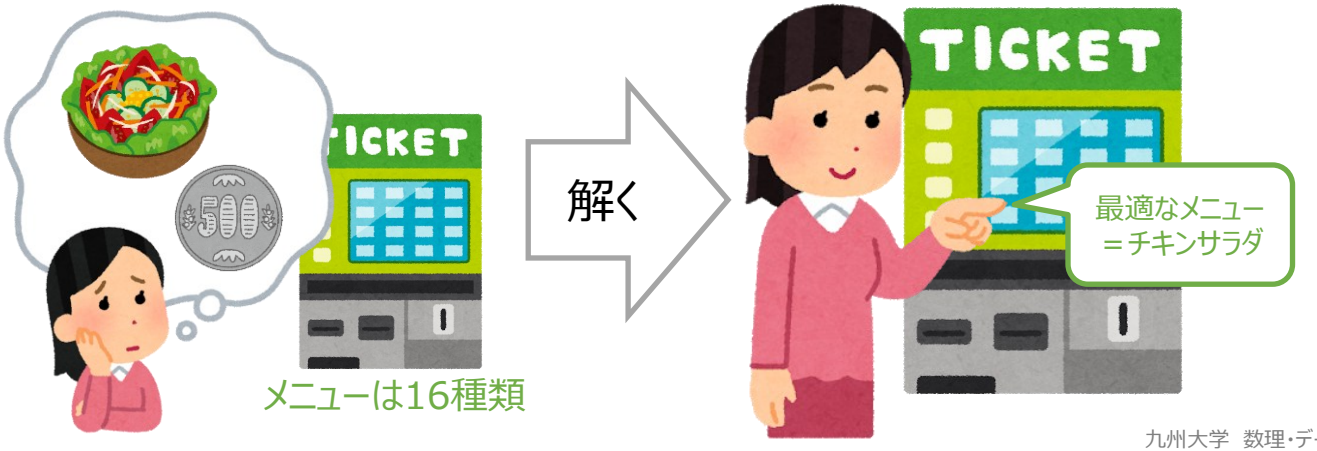


- 子供も大人も，日常的に（無意識に）やっている！
  - 最適化結果のうまい・ヘタはありますが…

最適化の三要素

- 制御変数
  - メニューの種類
- 目的関数
  - なるべく野菜をたくさん食べたい
- 制約条件
  - 昨日食べたのとは違うもので，500円以下

難しそうですが，
皆さんがやっている
最適化問題も
これらで表現できます



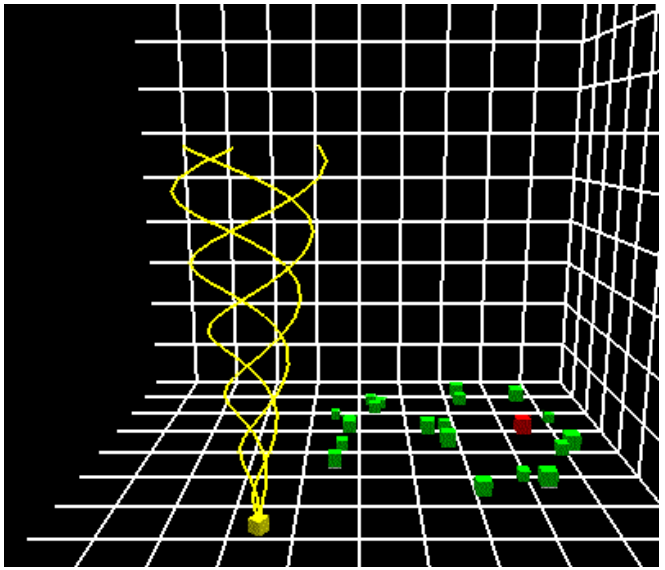
データ分析の関連話題：
シミュレーションとデータ同化

「シミュレーション」とは

- 和訳すれば「模倣」もしくは「模擬」
  - similar (似てる)と同じ語源

- 科学的目的のために「何らかの物理的法則」と「コンピュータ」を使って実際と似たような状況を作り出すこと

- 例
  - 過去に例のない大雨や大地震による被害を予想
  - 天気予報のために数時間後の気圧配置を予想
  - 人口密度とウイルスの拡散・消滅の関係を推定



NASAによる竜巻シミュレーション
Wikipedia “シミュレーション”

シミュレーションの意義：
以下のようなことでも，実際に起こったかのように観察できる

- まだ起こったことがないこと
  - ex. 地球の気温が40度になった時の海水面変化
- めったに起きないこと
  - ex. 津波や大地震が起きたときの状況
- 倫理的問題やコストの問題で実験できないこと
  - ex. 10年間ジャンクフードを食べ続けたときの体調変化
- 大規模過ぎてすべてを観察できないもの
  - ex. 地球全体の大気の状態

GenGan@Wikipedia 地球シミュレータ

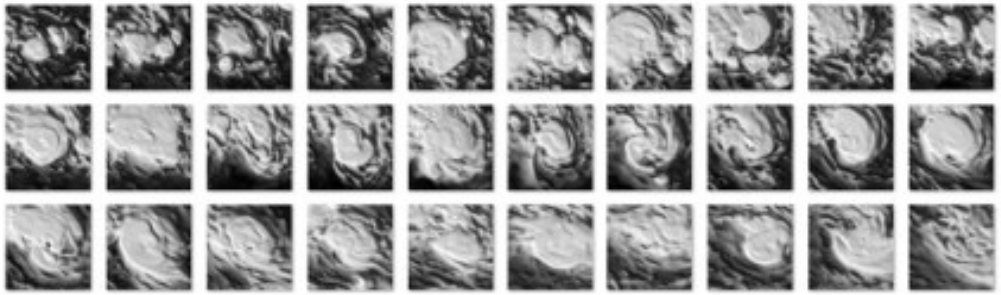
シミュレーションはデータ解析にも役に立つ (1/2)

- シミュレーションにより（人工的ではあるが）無尽蔵にデータが作れる

理論 → シミュレーション → 人工データ → 本物データの代理として分析

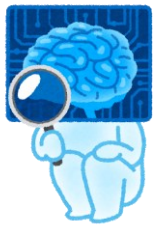
- もしシミュレーションが十分に正確なのであれば，実際のデータの代理として使える
- 例えば，「台風の発生理論に基づくシミュレーション」で生成したデータを分析して，台風発生のメカニズム解明や予測に利用する

シミュレーションで作成した台風の赤ちゃん



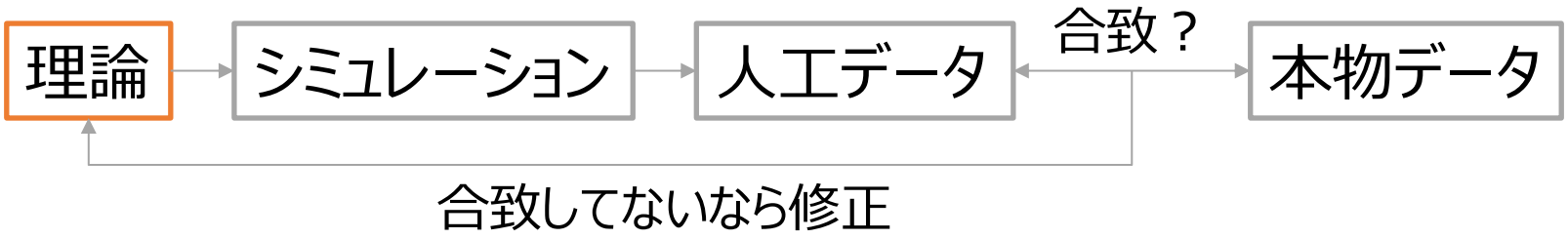
[Matsuoka+, Progress in Earth and Planetary Science, 2018]

AIによる台風の予測に利用



シミュレーションはデータ解析にも役に立つ (2/2)

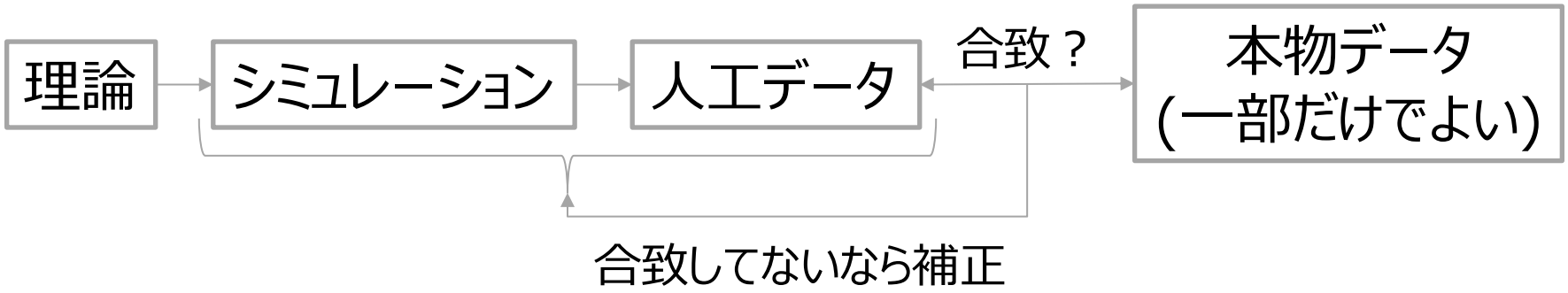
- 実データを用いて，シミュレーションやそのベースとなっている物理的理論の妥当性を検証



- ある「物理的理論」に基づいてシミュレーションを行う
- その結果が実際データとどの程度合致しているかを検証
- 合致しているのなら，もとの理論は正しい
- 合致してないのなら，もとの理論を修正

シミュレーションとデータ同化 (1/2)

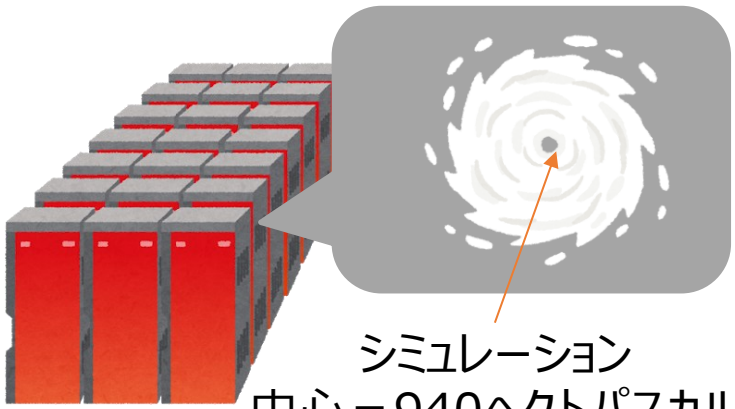
- データ同化＝実データを用いてシミュレーション結果を補正



- 皆さんも，頭の中で予想していた(シミュレーション)していた事態から，現実が一部でもズレていたら，予想を全体的に補正しますよね？

シミュレーションとデータ同化 (2/2)

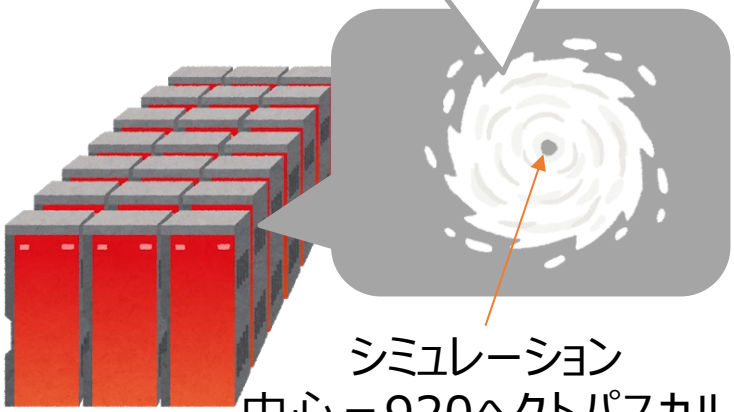
- 例：台風の中心の気圧だけで，全体を補正する



シミュレーション
中心＝940ヘクトパスカル

補正

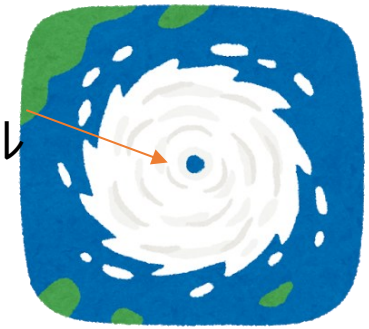
周辺部の気圧も
より正しい値に



シミュレーション
中心＝920ヘクトパスカル

ずれ

実データ
中心＝920ヘクトパスカル



結果的に
シミュレーション結果を
高精度化できる

因果関係を正しく知りたい人へ：因果推論と効果検証の基礎

ダイエットでやせたのは，本当にそのサプリのおかげ？

因果推論と効果検証

- 因果推論
  - 「ある原因がある結果を引き起こしているのかどうか」を明らかにする

- 効果検証
  - 「意図的に与えた原因」の結果への影響（効果）を明らかにする
  - 例：「サプリを飲んだら」（意図的に与えた原因）→「体重が減った」（結果）

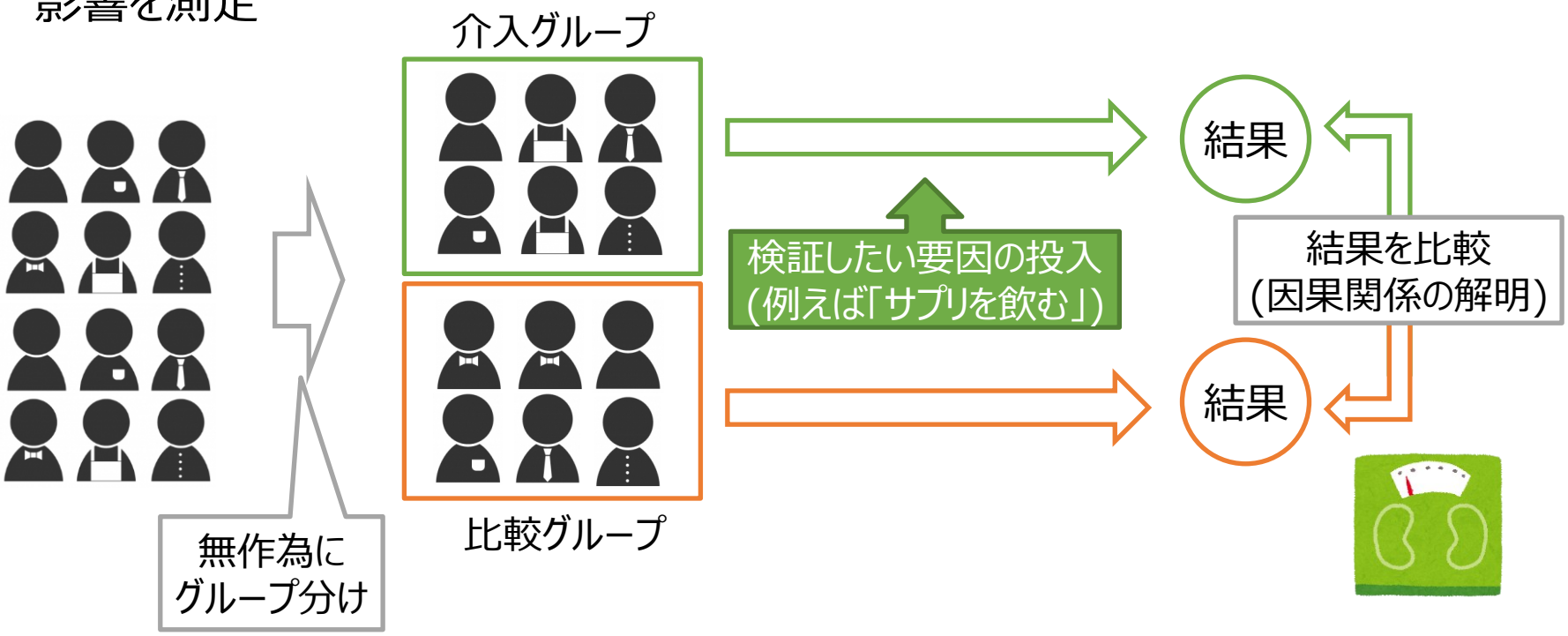
- 基本的考え方
  - 原因の有無で結果がどう変わるかをチェック

- 効果検証の方法
  - 原因の有無を積極的に作る方法　→　ランダム化比較試験とA/Bテスト
  - 勝手にできた「原因の有無」を利用する方法　→　自然実験

「原因の有無を積極的に作る」効果検証法(1)
ランダム化比較試験
Randomized Controlled Trial (RCT)

- 検証したい要因以外は公平になるように，対象の母集団を無作為にグループに分け，その検証したい要因の影響や効果を明らかにするための比較方法

- 具体的には「介入グループ」と「比較グループ」の2種類に分け，その試験的要素の影響を測定



ランダム化比較試験の例

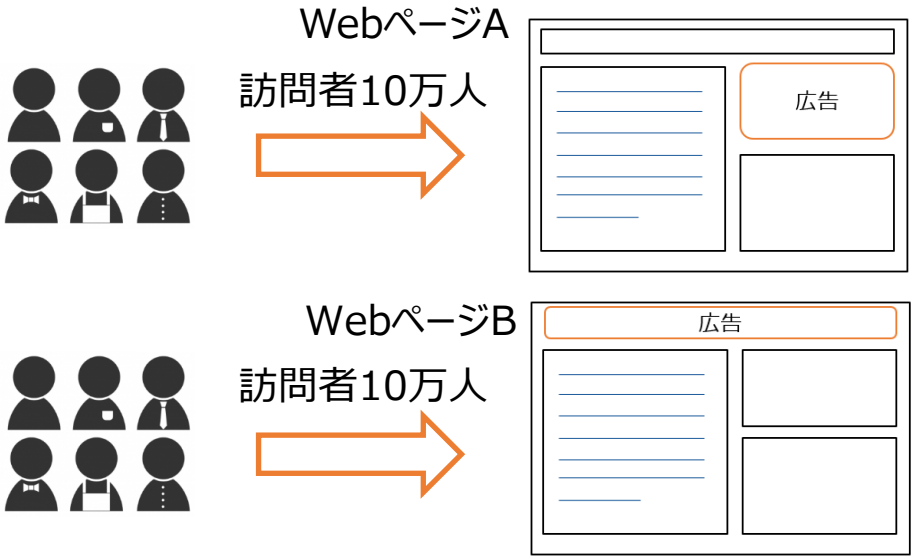
- 電力価格を上げると本当に節電につながるのか？
  - 参加者：北九州市内の一般参加世帯
  - 介入群：電力の需給が特にひっ迫する数時間の間，節電を促すための価格上昇を経験
  - 明らかにしたい因果関係：電力消費量に差が出れば，電力価格の上昇が電力消費量に影響を及ぼす
- 実験結果から
  - 電力価格の上昇は節電を促すという因果関係が分かった
  - 料金を上げるほど，価格の上昇に応じて節電が進む

参考：依田高典 田中誠 伊藤公一朗，「スマートグリッド・エコノミクス　フィールド実験・行動経済学・ビッグデータが拓くエビデンス政策」

「原因の有無を積極的に作る」効果検証法(2) A/Bテスト

- 2通り(以上)のパターンを用意し，どちらがより効果が高い成果が出るのかを検証する方法
  - インターネットのマーケティング分野で主に使われる
  - ランダム化比較試験の考え方を基礎にしている
  - オバマ氏も大統領選挙でより多くの支援者を獲得するために活用した手法

- 例) どのように広告を掲載すると（原因），クリック率が上がる？（結果）



どちらの広告配置がより多くの人にクリックされたか(広告商品の売り上げに貢献したか)を調査

ランダム化比較試験（A/Bテスト）の問題点と，解決策としての「自然実験」

- 実験に必要となる費用や労力などが膨大
- 各グループに十分な数の調査対象が必要
- 状況によっては，ランダム化比較試験を実施できない
  - 医療費の自己負担額を変化させると，医療サービスの利用頻度にどのような影響があるか
  - 所得税を低くするすると，その国(地域)に移住する人は増えるのか

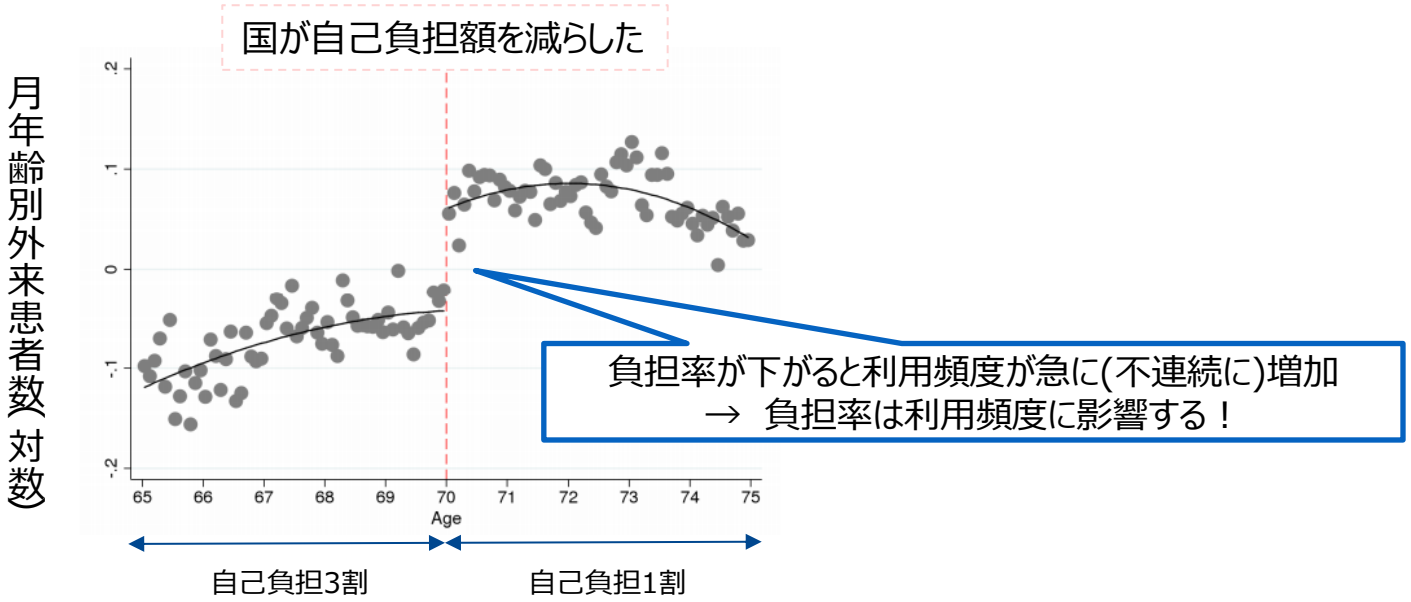
こんな実験は
倫理的にも社会的にも
難しい

- **自然実験**を利用
  - 自然実験＝「自然に（＝勝手に），比較実験と同じような状況ができた」
  - その状況を「うまく」見つけて使って，効果検証する

「自然実験」による効果検証の例：RDデザイン

Regression Discontinuity (RD) design

- 世の中に(調査とは無関係に)発生した「不連続」を用いた効果検証
  - 例：「国が医療費の自己負担率を下げた」ことを利用して，
  - 医療費負担率（原因）と医療サービス利用頻度（結果）の関係を調査



Hitoshi Shigeoka, 2014. "The Effect of Patient Cost Sharing on Utilization, Health, and Risk Protection," American Economic Review, American Economic Association, vol. 104(7), pages 2152-84